Predicting the popularity of online articles based on user comments

Alexandru Tatar
UPMC Sorbonne Universités
alexandru-florin.tatar@lip6.fr

Panayotis Antoniadis
UPMC Sorbonne Universités
panayotis.antoniadis@lip6.fr

Marcelo Dias de Amorim
UPMC Sorbonne Universités
marcelo.amorim@lip6.fr

Jérémie Leguay
Thales Communications
jeremie.leguay@gmail.com

Arnaud Limbourg
20Minutes
alimbourg@20Minutes.fr

Serge Fdida
UPMC Sorbonne Universités
serge.fdida@lip6.fr

ABSTRACT

Understanding user participation is fundamental in anticipating the popularity of online content. In this paper, we explore how the number of users' comments during a short observation period after publication can be used to predict the expected popularity of articles published by a countrywide online newspaper. We evaluate a simple linear prediction model on a real dataset of hundreds of thousands of articles and several millions of comments collected over a period of four years. Analyzing the accuracy of our proposed model for different values of its basic parameters we provide valuable insights on the potentials and limitations for predicting content popularity based on early user activity.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Computer Applications**]: Social And Behavioral Sciences

General Terms

Human Factors, Experimentation

Keywords

Online social networks, social media, content popularity, prediction, user behavior

1. INTRODUCTION

Online news platforms have become an important and very attractive source of information. They allow easy access to the latest news along with the integration of social media techniques that incites online readers to interact with the platforms through comments, votes, and online social links.

It has been observed that online content has a disproportionate distribution of user interest, where a small number of items generate a lot of attention while the great majority is barely noticed. An example is the unequal distribution of popularity for news stories observed by Huberman et al. on the Digg portal [1], [2]. They have found that from the 7% all of the submissions that gather sufficient votes to be promoted on the front page (the rest of the stories receive for the most part only 1 vote), 30% receive more than 1000 votes.

Given such a high variability of attention, publishers can significantly benefit by *predicting the expected popularity of news articles*. The reasons are manifold. First of all, it can be used as a powerful tool in online marketing for more efficient advertising placement. Also, both online readers and news platforms that deal with a huge amount of information can use predictions for content filtering of different types. For example, one can integrate such predictions in the way content is organized on the website and thus improve the user experience while browsing.

In a quite different context, predicting content popularity would be of major interest in opportunistic communications among mobile devices. Indeed, the high adoption rate of powerful mobile devices has brought closer the vision of mobile social networking and numerous applications which enable the direct content exchange between users in proximity are being developed [3, 4, 5]. However, in such resource-constrained environments (e.g., limited contact time and/or bandwidth between devices, storage constraints), predictions can be used in conjunction with ranking methods for content replication or replacement techniques.

There are many different metrics that can express the popularity of content when considering different types of user activity and feedback, such as the number of views, comments, the number of votes or other types of rating values, and the number of times the information has been shared through online social networks or emails. Understanding the predictive characteristics of each one of these metrics is a difficult task and consists in finding the appropriate attributes that describe the evolution over time. These attributes can include some generic content characteristics (such as topic, author, publication hour, etc.) or early information of the popularity metric under observation.

Indeed, the use of the recorded activity as a metric for predicting content popularity is considered more and more in recent research, which demonstrates that we can make

WIMS'11, May 25-27, 2011, Sogndal, Norway.

realistic predictions on videos [1], news stories [6], podcast channels [7], and even movie box office revenues [8] using different aspects of user activity and different prediction methods.

However, there is no clear evidence that there is always a prediction model that could be applied for every possible scenario nor that the creation of a generic prediction model is a feasible objective. The main reason is that predictive methods can be highly influenced by the type of data set and more particularly by the site's framework, the size of the data set, or other important details [6]. Moreover, research studies can only be based on publicly available information which is not always complete and/or accurate.

For example, the number of views could be considered in principle an adequate measure of popularity. This is so, especially since it has been shown in [9] that users that actually contribute in social media platforms (e.g., weblogs or wikis) represent a proportion of the users that only read the content. However, the actual number of views are not always available to external observers. And when available, their number is mostly computed as the number of times a specific web page is requested, which could be due to web crawlers, search engines and can be very easily manipulated both by content publishers (to boost the perceived popularity of their content) and content creators for the same reason.

Similarly, rating information could be also manipulated especially in cases of low activity or could even be subject to reciprocity strategies by the users[10]. The same holds for comments and any other popularity metric especially when users realize that it is used as an input to promotion techniques based or not on predictions.

Our objective is not to provide a final solution to this complicated problem but to shed some more light on the factors that affect the prediction quality of online content. We do this by adding to the existing literature our prediction analysis of a rather interesting case study. More specifically, we have the privilege of having access to a huge dataset of a major countrywide online news publisher composed of hundreds of thousands of articles and millions of comments over a period of 4 years. To the best of our knowledge this is the first kind of research conducted on data set so complete.

In this paper, we present a simple, yet efficient methodology to evaluate the expected popularity of news articles after their publication. In our context, the only metric that indicates the popularity of an article proposed by the site is the number of comments and there is no other rating functionality offered to the users. As a result we consider the number of comments as an implicit evaluator of the interest generated by an article. We use the number of comments received by an article early after its publication as the main indication of the final interest that it will generate.

We also study the possible benefits of the specialization of the prediction model with different article characteristics such as the publication hour and the category of news it belongs to. Although one would expect that specialization should increase the accuracy of the prediction for such important article characteristics (especially given the fact that our data set is big enough to provide a large training set for each specialization category) our results indicate that this does not hold in practice.

Interestingly, there are many cases in which we require only a short period of observation to deliver accurate prediction levels even if we use a single generic prediction function calculated with data belonging to a short history of activity. This means that there is a potential benefit from applying prediction models like the one we study in this paper in real scenarios.

As a summary, this paper has the following contributions:

- We characterize the dynamics of articles and comments and investigate how these characteristics evolve over the years in a countrywide participative news media site.
- We analyze in depth the use and trade-offs of a simple linear regression model as a prediction method for different training and test sets, prediction models, observation periods, and accuracy metrics.
- We make the non-intuitive observation that specialized prediction functions and larger datasets do not increase significantly the achieved accuracy and demonstrate that in many cases a simple and generic prediction model is a sufficient solution to predict the expected volume of comments for meaningful observation periods.

The rest of the paper is organized as follows. In Section 2, we briefly describe the data sets used in our analysis. In Section 3, we provide some global statistics on the publication of articles and comments, while in Sections 4 and 5 we describe and evaluate a linear prediction model. We present the related work in Section 6 and conclude the paper by suggesting future research directions in Section 7.

2. DATASETS

We have conducted our analysis on several datasets obtained from 20Minutes web site. 20Minutes is a free daily newspaper published in the main cities in France. According to a survey conducted by Audipresse in 2009,[1] it is the most read daily journal in France with an average number of 2,675,000 readers per day. 20Minutes also comes with an online edition of the newspaper. The news site reaches more than 4 million unique visitors and 73 million site visits per month.[2].This makes 20Minutes the fourth visited online press site in France. 20Minutes's targeted audience consists of active readers with an age between 15 and 40 years old. The content of the site is news oriented, starting with the main articles from the printed version and being periodically updated with the latest news.

Throughout our analysis we have employed two datasets consisting of dump files of the articles and the associated comments. The time period covered by the datasets is more than 4 years, starting from February 2006 until June 2010. The size of the datasets is also significant, counting 338,394 articles and 2,666,876 comments. For some of the articles, comments have been disabled by the moderator within some time after their publication. This is typically done to prevent some forms of abuse. As a consequence, we do not take into consideration these articles in our analyses. After removing these articles(23 % of all articles), our data set contains approximately 260,000 articles and 2,500,000 comments. The most consistent part of our research is based on

[1] http://www.audipresse.fr

[2] http://www.mediametrie.fr

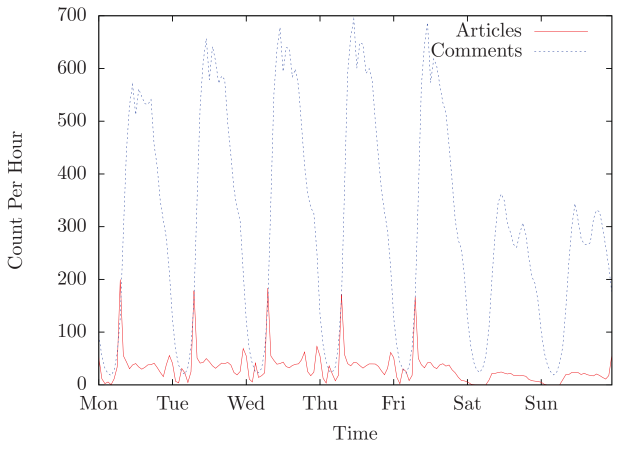

Figure 1: The number of articles and comments posted per hourly cycles. On the y-axis, we count the average number of comments and articles published per hour. The peaks corresponds to 7→8 a.m. period for articles and 11→12 p.m. for comments.

comments but, when needed, we also make use of essential information about the articles such as the creation time or the category the article belongs to.

3. GLOBAL STATISTICS

We begin our analysis by providing general statistics on the dynamics of comments and article publication. Our dataset indicates that the average number of articles per year was 63,000 with the observation that, starting from 2007, the number of published articles has decreased by 15% per year. Nevertheless, the number of comments per year has continued to increase. From 2006 to 2008 the number of comments has doubled each year reaching almost 800,000 in 2008 and stabilizing around this value.

Our data set confirms previous observations of the circadian pattern of content generation [7, 11]. Figure 1 shows the number of published articles and the number of comments submitted by 20Minutes readers, on daily and weekly basis.[3] We can observe that the majority of articles are published between 7 a.m. and 8 a.m. and that user comments are mostly received in the 11→12 p.m. period. Daily variations can also be deducted from this graph. Readers are twice more active during the working days compared to the weekend, nevertheless with an important contribution if we consider the number of published articles (fewer during weekends).

It is interesting to compare how content generation has evolved during the 4 years. On one hand, we observe that the proportion of comments per hour has insignificant variations despite the fact that the site's publishing strategy has undergone significant modifications (Figure 2). On the other hand, the publication of online articles has changed during the years and we observe that the site is now publishing its articles more evenly during the day.

Table 1 presents interesting statistics on different categories of articles. We compare the percent of commented articles, the average (and median) number of comments per article, and the median time between the publication of an article and its first and last comment. For space constraints, we present here only 7 categories, the most important ones in terms of number articles and comments. We observe that the percent of commented articles differs per category. For example, articles belonging to *Debates* are extremely commented while *Economy* is on the other extreme, with only 14% of commented articles. As a general observation, the overall ratio of commented articles is 38%, compared to 23% found in the literature [12]. We also observe that 20Minutes articles receive their first comment earlier. The median elapsed time is 1.5 hours compared to 6.7 hours on news [12] and 2 hours for blogs [11]. Our data also suggest that besides *Debates* section, which is opened for long discussions, the articles receive their last comment on average within one day.

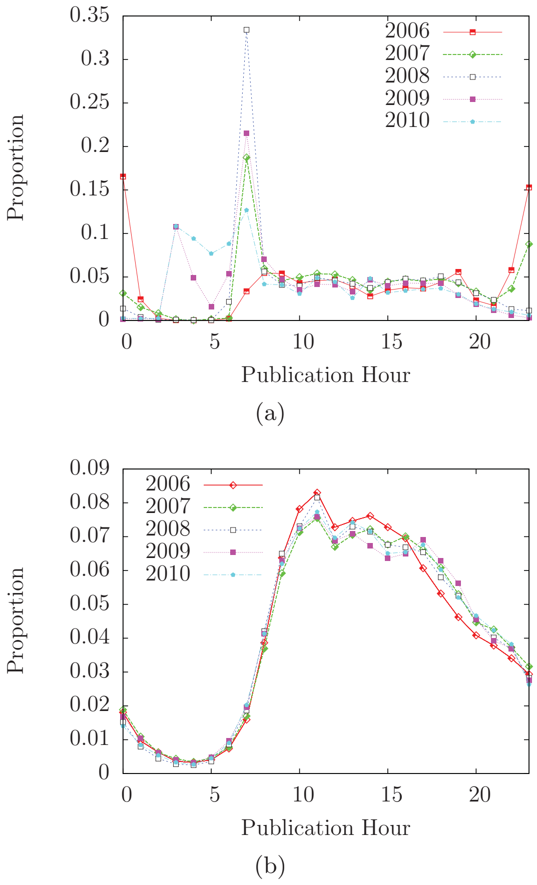

Figure 2: The proportion of articles and comments that were published per hour. The x-axis represents a 24-clock interval. Figure (a) shows the ratio of articles per publication hour while Figure (b) follows the ratio of comments.

4. PREDICTIONS

4.1 Methodology

Our prediction mechanism is based on a linear regression

[3] We refer to the average value per day covering the 4 years period.

Table 1: Statistics for articles belonging to different categories. The first and last comment times represent the median values and are expressed in hours.

	Articles	Comments					
Category	Total (Commented)	Total	Avg.	STD.	Med	first[hrs]	last[hrs]
France	36,106 (60%)	951,857	43	88	13	0.8	25
World	28,108 (41%)	262,899	22	43	8	1.1	21
Sports	24,667 (45%)	191,136	17	36	7	1.3	20
Economy	29,990 (14%)	108,086	24	58	3	1.3	17
Paris	8,377 (42%)	55,477	15	38	3	8	32
Politics	12,943 (58%)	251,331	33	77	4	0.8	17
Debates	840 (97%)	144,889	177	300	89	0.3	265

model that uses a single variable:

$$Y = f(X) = \alpha + \beta X \quad (1)$$

In our scenario, Y is the predicted variable and consists in the total number of comments that an article has received and X is the independent variable that expresses the number of comments an article has received some period of time after its publication. We will thereafter refer to the time difference between the prediction moment and publication time as *observation period*. X will therefore express the number of comments received by an article during an observation period. We measure the prediction characteristics of our model using the adjusted coefficient of determination, R^2_{adj} [13]. This statistical parameter explains the measure of variation of the independent variable that is predicted by the dependent one. R^2_{adj} is a good measure of studying if a regression model can predicting future outcomes. Throughout our analysis we evaluate the predictive characteristics of articles for different observation periods. The promptness of a prediction is a crucial factor for a prediction mechanism. We will therefore analyze our prediction power for observation periods of less than 24 hours, the median duration in which articles received their last comment.

4.2 Article characteristics

Our data set suggests that the predictive characteristics of articles may differ significantly per year. As can be seen from Figure 3 the articles published in 2006 and 2007 present poor predictive characteristics compared to 2008-2010 data, which surprisingly presents similar characteristics. We believe that the weak results found for 2006 and 2007 data are due to the low interest shown by online readers, their contribution being significantly lower compared to the following years. Figure 3 also indicates that for the articles published in 2008-2010, an observation period of 24 hours is a very strong indication of the total number of comments that articles will receive.

We have observed in Section 3 that articles may receive different attention from the readers based on their category. In addition, Figure 2 reveals that the volume of articles and comments published during the day is highly variable. These observations suggest that a separation and analysis of articles per submission characteristics (publication hour and category) could expose some interesting perspectives on how to better predict the popularity of news articles. Indeed, by examining Figure 4 we observe that some sections present better predictive characteristics that others.

A similar observation can be made by grouping articles based on the publication hour. We can observe in Figure 5 that the predictive characteristics can be very different depending on the publication periods. Our data sets suggests that articles published between 0 and 5 a.m. present low predictive results even for an observation period of 24 hours, whereas those published between 6 and 11 a.m. interval show good results after just 5 hours of observation.

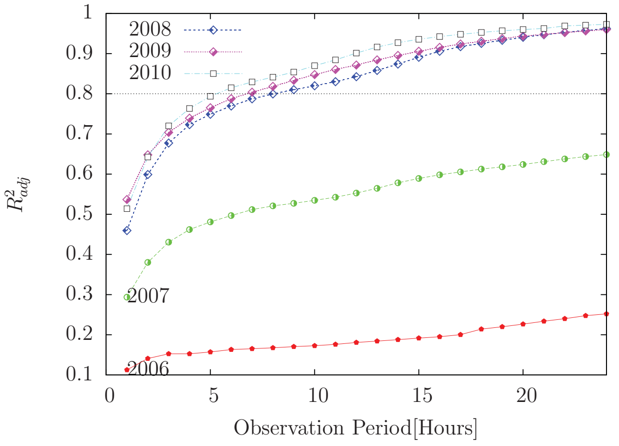

Figure 3: The R^2_{adj} for different observation periods. The articles are grouped per publication year.

4.3 Minimum observation period

It is intuitive to assume that by increasing the observation period we achieve better prediction levels. But in a practical situation we should not neglect the promptness of a prediction. A system that integrates a prediction mechanism should provide estimations as soon as possible, i.e., within the shortest possible observation period. For this scenario we are interested in how many hours a system would need to observe in order to make predictions with a minimum accuracy. We consider a minimum prediction accuracy threshold $R^2_{adj} \geq 0.8$ (which suggests a strong predictive relationship) and we record then the first observation period that exceeds this value.

Based on our previous observations we group and analyze articles per year, category(those presented in Table 1) and publication hour. We obtain more specialized groups of articles for which we compute the minimum observation period that attains an $R^2_{adj} \geq 0.8$. The clustered histogram shown in Figure 6 compares the proportion of subsets of articles

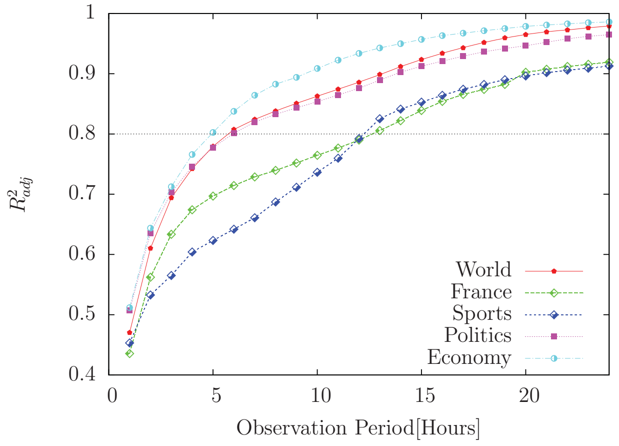

Figure 4: The R^2_{adj} for different observation periods. The articles are grouped per category.

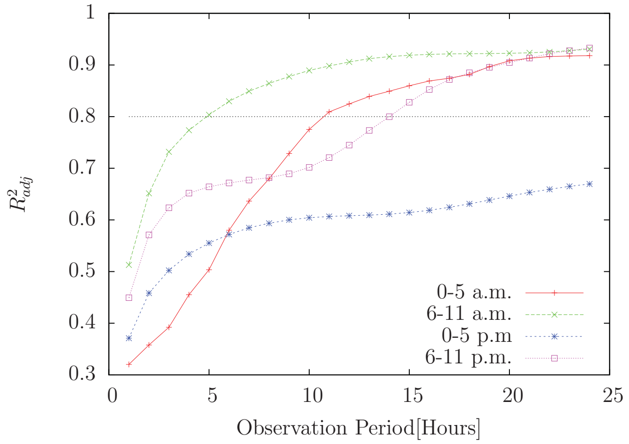

Figure 5: The R^2_{adj} for different observation periods. The articles are grouped per publication hours.

that verify a certain observation period. We did not include articles published in 2010, as this year was not fully covered in our data set. Surprisingly, we have found that an observation period of only 3 hours is often enough to suggest a strong prediction. The results presented in Figure 3 suggest that the articles published in 2007 have moderate predictive characteristics. Accordingly, we can observe in Figure 6 that a high number of 2007 articles need an observation period of more than 24 hours for an accurate prediction.

5. EVALUATION

In order to evaluate the prediction power of our model and its dependence on articles characteristics, we separate our data into training and the test set using the holdout method of separation[14]. Using this method, we divide the dataset in two disjoint parts, training and test set. The prediction model is obtained from the training set and its prediction accuracy is evaluated on the test set.

For this phase of our evaluation, we have considered all articles published in 2010 (21,000 articles) for the test set[4] and use as training set the rest of the articles.

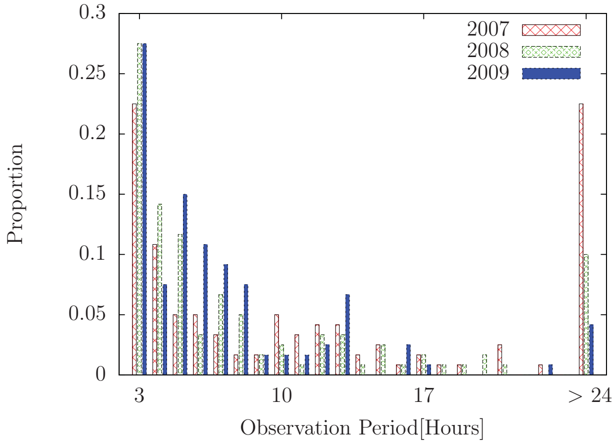

Figure 6: The proportion of sub of articles for which we obtain an adjusted $R^2_{adj} \geq 0.8$. Each observation period is calculated for a subsets of articles published at a specific hour and belonging to one of the categories presented in Table 1. Articles are also separated per year.

5.1 Prediction accuracy

We evaluate the prediction strength of our model using two evaluation criteria. First, the Root Mean Squared Error (RMSE) as defined in Equation 2.

$$\text{RMSE} = \sqrt{\frac{\sum_{1\leq i\leq n}(f(x_i) - y_i)^2}{n}}. \tag{2}$$

Second, the Kendall rank correlation coefficient, which is defined as follows. If we consider two samples, x and y, each of size n and the total number of possible pairings of x with y observations ($n(n-1)/2$), S is the difference between the number of concordant (ordered in the same way, n_c) and discordant (ordered differently, n_d) pairs.

Then the Kendall coefficient τ is related to S by:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \tag{3}$$

These two evaluation measures allow us to examine our model from different point of views. The RMSE expresses the accuracy of our model in terms of average number of comments (when comparing the real and predictive values) while Kendall coefficient is used to compare the level of similarity between a real and predictive ordering of articles in terms of total number of comments. Clearly, both are conservative evaluation criteria since in practice even a simple classification of articles into a few popularity classes could be sufficient for making informed decisions in terms of content placement or filtering (see also Section 7). Therefore,

[4] We exclude the articles received in the last two months of 2010. These articles may still be active and therefore we cannot estimate with confidence the total number of comments received in their lifetime.

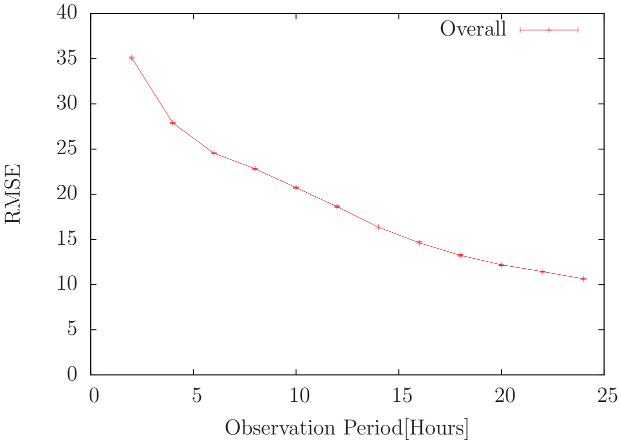

Figure 7: Prediction strength measured by the RMSE for different observation periods from 2 to 24 hours.

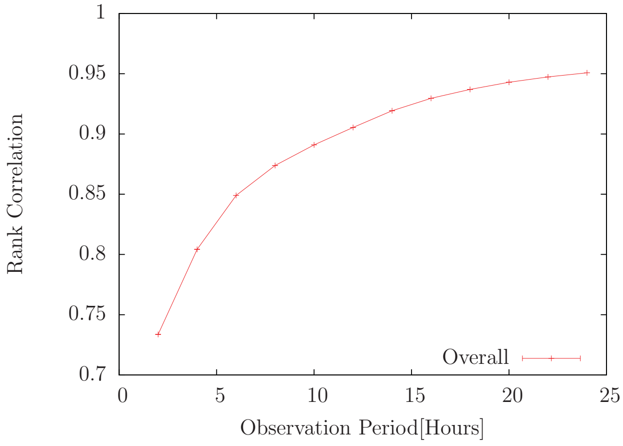

Figure 8: Prediction model strength measured by the Kendall rank correlation for different observation periods from 2 to 24 hours.

the results presented in the following could be considered as worst cases for many practical uses of our prediction model.

5.2 Observation Period

Our linear model is based on the analysis of early users' interest in articles and it is of great interest of achieving an accurate prediction in a very short observation period. We therefore investigate how the observation period influences the accuracy of the prediction. We analyze this accuracy using both the average error of comments(RMSE) and the ranking of articles(Kendall rank correlation).

Figures 7 and 8 depict the achieved accuracy as quantified with the above metrics when we use an *overall* prediction function. This function was obtained by taking into consideration the entire training set (2006-2009 articles). As we observe, after only 5 hours of observation we achieve a satisfactory prediction accuracy especially if we recall that our evaluation metrics are rather conservative. We also can notice that for one day of observation we could accurately rank articles based on the total number of comments that they will receive during their lifetime.

5.3 The role of history

In our previous example we have obtained our regression model using a training set of almost 4 years. However, in practical scenario, one may not have access to such a big dataset. In addition, as Figure 3 suggests, data may present particularities per year that could influence the prediction function. We therefore vary the size of training set starting with articles published in December 2009 and incrementally increasing the training period up to 4 years using a 3 months granularity. This separation method allows us to observe how the training set size influences the prediction accuracy. Interestingly, in our dataset even the smallest selected period of 3 months is adequate to achieve the best possible predictions(Figures 9). This means that one does not need a huge amount of data to reach a satisfactory level, which increases significantly the possible applications of simple prediction models like the one proposed in this paper.

So, the size of our dataset did not prove useful for improving the prediction accuracy. However, it allows us to experiment with different specialized versions of our model which considers different subsets of articles as we discuss below.

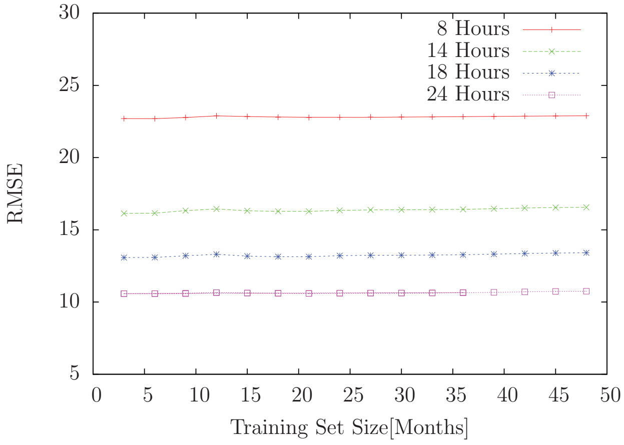

Figure 9: The RMSE as a function of the training set size for different observation periods.

5.4 The role of specialization

Our earlier observations on the predictive characteristics of articles suggested that a clustering of articles per different characteristics (publication hour or category) could be beneficial for the prediction accuracy. As a result we have compared the prediction accuracy of an overall prediction model with various specialized models that create separate prediction functions for different categories of articles, more specifically belonging to different sections and having been published at different hours.

Figure 10 compares the prediction accuracy of the 3 models(*overall*, *per section* and *per hour*) over the same test set composed of 2010 articles. We can observe that the *overall*

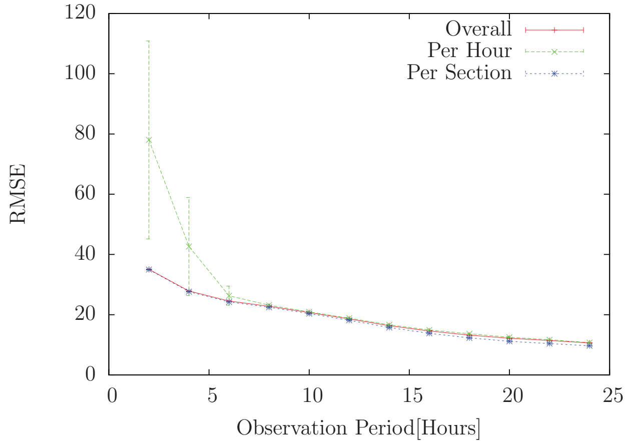

Figure 10: The prediction strength of the *overall* model compared to more specialized ones: *per section* and *per hour*. Each point represents the mean and standard deviation of the RMSE for all training sets.

and *per section* models present similar and stable characteristics in terms of RMSE. In addition it can been seen that a classification per publication hour is not recommended for observation periods less that 8 hours. These results suggest that the overall prediction model is adequate and although there are scenarios in which the specialized functions perform slightly better, the extra gain does not justify the complexity for producing them. This conclusion is reinforced when we consider the Kendall ranking coefficient in which cases the differences in accuracy achieved between the overall and specialized functions is zero for all scenarios.

6. RELATED WORK

Our work covers two main research domains: the dynamics of content generation and prediction methods based user generated content. A first set of analysis on the commenting characteristics found in the blogosphere was provided by Mishne and Glance [11]. They have studied the relationship between the weblog comments and the posts and have shown that comments are a good indication of popularity for a weblog. Important remarks on the weekly and daily generation of comments and articles, from a technology-news website can also be found in [15]. They have also observed a daily and weekly pattern of comment generation and found that the daily activity reaches its maximum at around 1pm.

The closest to our work is the analysis of articles and comments from several Dutch online news platforms by Tsagkias et al. [7, 12], which has revealed important traits on the generation of articles and comments. They have found patterns in the distribution of comments throughout the day and demonstrated that these characteristics depend on the specific online news source. Our analysis is conducted on a similar volume of articles but differs in number of comments, survey period, article categories or commenting facilities. We also show that comments have a strong prediction power and we analyze in more depth the effect of various system parameters on the achieved accuracy.

Predicting the popularity of online content using different characteristics of user interactions has shown its efficiency in many fields. There are a great number of prediction methods and their selection depends on the purpose of the prediction and on the type of data set analyzed. Tsagkias et al. used different podcast features to achieve two different prediction tasks, classification and ordering of podcasts [7]. An interesting prediction outcome can be found in the work of Asur and Huberman [8]. They investigate the prediction accuracy for movie box office revenues using the chatter from Twitter. They have showed that there is a strong correlation between the number of comments related to a movie and its box-office revenues. Their results are more accurate than those of the Hollywood Stock Exchange. In our paper we have used a similar prediction method but on a very different and much bigger data set.

Another approach on predicting the popularity of news (on Digg portal) or videos (on Youtube) using early observations, can be found in [1]. Using a simple logarithmical prediction method and the number of votes, they have measured the correlation between the interest shown in the beginning of the publication and the final popularity of the content. We have observed similar characteristics for stories saturation, although the data sets and methods differ considerably. Tsagkias et al. have also analyzed the possibility of predicting the volume of comments for articles, prior to the their publication, using a non-linear prediction mechanism on semantical and textual features [16]. They have showed that semantic and textual features can be good predictors in identifying and classifying articles that will receive a high number of comments. Such analysis could complement prediction methods as the ones analyzed in this paper.

7. SUMMARY AND OUTLOOK

We have shown in this paper how a simple linear regression model can help predict the volume of comments for news articles and build expected popularity rankings, which can inform content placement and filtering strategies. To this end, we have evaluated a simple and efficient prediction metric that is the number of comments observed within a certain time after the publication of an article.

Our analysis focuses on a single source of data that may present distinctive particularities compared to other daily news platforms, e.g. sites design, commenting method or the generation of articles. It is difficult to envisage how any of these parameters could influence the prediction process.

However, our first results include some interesting non-intuitive observations that could have some general applicability. More specifically, there are strong indications that specialized prediction functions and a long history do not necessarily increase the accuracy of the achieved prediction. To reinforce this interesting observation we plan in our future work to experiment with more sophisticated prediction strategies (e.g., non-linear models, inclusion of more dynamic parameters like the time before the first comment and others) on the same data set. Then assuming that indeed the history of observations does not improve significantly the prediction accuracy we can extend our study to various different news sites and compare the performance of our prediction model.

Such a comparative analysis will be of great importance for assessing the applicability of prediction models in scenarios for which their potential value is of critical importance. More specifically, we wish to study the feasibility of implementing a news distribution system on top of a mobile social network, building on a prediction mechanism to optimize resource utilization. For example, given the inherent constrained nature of such networks, it would be interesting to evaluate our prediction strategies for content prioritization (e.g., which articles should be removed from the system when the maximum storage capacity is reached or when there is a chance of content replication during a contact between devices).

The challenge in this context is the fact that mobile social networks are created ad hoc between diverse types of people and in diverse environments. Relying on a generic prediction model to optimize the distribution of content in this context would be meaningful only if there is evidence that this model has good performance in a wide variety of scenarios. We believe that in this paper we made a first step toward answering this and other interesting questions related to the predictability of content popularity in social media systems.

8. ACKNOWLEDGEMENTS

This work was partially supported by the ANR project Crowd under contract ANR-08-VERS-006 [17].

9. REFERENCES

[1] G. Szabo and B. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[2] F. Wu and B. Huberman, "Novelty and collective attention," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, p. 17599, 2007.

[3] N. Vallina-Rodriguez, P. Hui, and J. Crowcroft, "Has anyone seen my goose? social network services in developing regions," in *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, 2009.

[4] N. Eagle and A. Pentland, "Social serendipity: Mobilizing social software," *IEEE Pervasive Computing*, pp. 28–34, 2005.

[5] P. Persson, J. Blom, and Y. Jung, "Digidress: A field trial of an expressive social proximity application," *UbiComp 2005: Ubiquitous Computing*, pp. 195–212, 2005.

[6] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 621–630.

[7] M. Tsagkias, M. Larson, and M. De Rijke, "Exploiting surface features for the prediction of podcast preference," *Advances in Information Retrieval*, pp. 473–484, 2009.

[8] S. Asur and B. Huberman, "Predicting the future with social media," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, ser. WI-IAT '10, 2010.

[9] J. Nielsen, "Participation inequality: Encouraging more users to contribute," $http : //www.useit.com/alertbox/participation_inequality.html$.

[10] J. G. Lee, P. Antoniadis, and K. Salamatian, "Faving Reciprocity in Content Sharing Communities (A Comparative Analysis of Flickr and Twitter)," in *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2010.

[11] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," in *WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, at WWW ': the 15th international conference on World Wide Web*, 2006.

[12] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," *Advances in Information Retrieval*, pp. 191–203, 2010.

[13] N. Draper and H. Smith, *Applied Regression Analysis*. Wiley-Interscience, 1998.

[14] V. K. Pang-Ning Tan, Michael Steinbach, *Introduction to Data Mining*. Addison-Wesley, 2005.

[15] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López, "Homogeneous temporal activity patterns in a large online communication space," *IADIS International Journal on WWW/INTERNET*, vol. 6, no. 1, pp. 61–76, 2008.

[16] M. Tsagkias, W. Weerkamp, and M. De Rijke, "Predicting the volume of comments on online news stories," in *Proceeding of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 1765–1768.

[17] "Crowd project," http://anr-crowd.lip6.fr/.